質問に対する回答を含むWebページの発見手法

高田　夏希† 大島　裕明† 小山　聡‡ 田中　克己†
京都大学大学院情報学研究科† 北海道大学大学院情報科学研究科‡

1. はじめに

QAサイトには，過去に投稿された質問と回答集合の組（以下QAコンテンツ）を閲覧することで情報を得るという利用方法がある．しかし，質問の内容によっては回答可能なQAサイトユーザが少ない場合や，質問に対し有益な情報を持つユーザが質問の存在に気付かずに回答情報が与えられない場合がある．そのためQAコンテンツの中には質問に対する回答が十分でないものが存在する．

そこで，我々はこれまで，QAサイトに存在するQAコンテンツに対し，既に与えられた回答とは異なる別解となるような回答をWeb上の別のリソースから発見してくる手法について取り組んできた．質問に対する回答をWebから取得する方法としては佐藤らの研究がある[1]．この研究はQAコンテンツへの補完ではなくユーザの入力した質問に対する回答取得が目的である点で本研究と異なっている．

本稿では，我々がこれまで提案してきた手法によって取得されたWebページ内に，対象の質問への回答となる情報が含まれているかどうかを判定する手法を提案する．

Webページに含まれる回答情報は，対象とするQAコンテンツに既に与えられた回答とは，ある一つの質問に対する答えであるという点で同種の情報であるといえる．そこで，QAコンテンツから質問と回答を表す特徴ベクトルを作り，それらを利用してWebページが回答情報を含むかを判定するためのスコアを計算する．

2. 質問に対する回答を含むページの判定

質問qを表す語集合KをクエリとしてWeb検索を行って得たページ集合Pの各要素p_nがqに対する回答情報を含むかどうか判定するためのスコア*Score*を，回答を表現する際に共通して現れる語の語集合Cを用いて計算する手法について述べる．qを表す語の抽出に関しては我々の研究[2]がある．本稿ではKは人手で与えるものとする．

本稿では，ある質問に対する回答集合において回答を表現する際に共通して現れる語の存在を仮定している．例えば，二日酔いの解消方法として「ウコンを飲む」，「大量の水を飲む」というものがあるがこれらの回答には“飲む”という語が共通して現れている．このような，回答に共通して現れる語を取得し，ページ内にその語が含まれるかを調べることでページに回答情報が含まれるか否かを判定することができると考えられる．

ここで，qに対する回答を述べる際に共通して現れる語の集合Cの取得方法について述べる．回答に共通して現れる語は，質問を表すような語集合と各回答a_mを表すような語集合を共に含む文書集合中に多く現れると考えられる．そこで，以下の手順でCを取得する．

1. 入力として与えられる回答集合Aの各要素a_mについてa_mを表すような語の集合Wa_mを取得する
2. KとWa_mの各要素をAND条件とするクエリでWeb検索を行い，それぞれで検索結果n件のスニペットを取得する
3. Aの要素数回2.を繰り返し，スニペット集合Sを得る
4. Sにおいて出現回数の多い語をCとして取得する

以下に，各手順について説明する．

[1.]Cを取得するためには各回答を表すような語を取得する必要がある．回答を表す語の取得には，馬らの研究における語の共起関係を用いた話題構造抽出手法[3]を用いる．馬らは主題度・内容度という指標を用いてテキストの主題語・内容語を抽出するということを行っている．

QAコンテンツにおける主題は質問文の内容とみなすことが出来るため，QAコンテンツの主題語は質問を表す語集合Kと考えられる．馬らの研究[3]では主題についてテキストが何を述べているかを表す語が内容語となっている．

QAコンテンツでは，質問を主題として各回答がそれぞれ主題について述べているものとみなせる．よって，馬らの研究[3]における内容語集合を本研究における回答を表す語集合に置き換えることができる．馬らの手法を用いて各回答を表す語をそれぞれ取得する．

[2.] 回答に共通して現れる語Cは，質問を表すような語集合と各回答を表すような語集合を共に含むWeb文書集合中，または両者を共に含むQAコンテンツ集合の回答中に多く現れると考えられる．よって，Cを得るために，Kと先ほどの1.で得られるWa_mをキーワードクエリとしてWeb検索エンジンまたは知恵袋検索を用いることで，KとWa_mを共に含む文書集合のスニペット集合またはQAコンテンツ集合の回答集合を得る．

[3.]2.をAの要素数回繰り返し，KとAのいずれかの要素を表す語を同時に含む文書集合のスニペット集合SまたはQAコンテンツ集合の回答集合A'を得る．

[4.]3.で得られたスニペット集合Sまたは回答集合

Finding Web Pages as Answers for a Question in a QA Site
† Natsuki TAKATA, Hiroaki OHSHIMA, Katsumi TANAKA(Kyoto University)
‡ Satoshi OYAMA(Hokkaido University)

A' において，出現回数（$tf(c)$とおく）の多い語 c を回答に共通して現れる語 C の要素とする．

ページ集合 P の各ページ p_n について，p_n に回答が含まれる可能性の高さを表すスコア $Score_{pn}$ の計算方法について説明する．p_n 内のリンク文字列以外のテキストに C の要素 c_i が含まれるかを調べ，含まれれば $Score_{pn}$ に $tf(c_i)$の値を足す．これを C のすべての要素について繰り返し，$Score_{pn}$ の値を計算する．ここでリンク文字列を対象外とするのは，そのページ自身に回答が現れるかを判定するためである．たとえば，「二日酔いに飲むと良いものはこちら」といったリンク文字列がある時，“飲む”という語は含まれているが，何を飲めば良いのかはそのページ内ではなくリンクをたどった先に書かれていることが考えられるためである．

3. 評価実験

ページ集合 P の各要素 p_n が質問に対する回答を含む可能性の指標 $Score_{pn}$ について，その妥当性について評価実験を行った．

2 節では，回答に共通して現れる語の集合 C の取得方法として

- (WA)質問を表す語集合 K（人手で与える）と各回答を表す語集合 Wa_m を用いて Web 検索結果のスニペットから取得
- (YA)K と Wa_m を用いて知恵袋検索結果の QA コンテンツの回答集合から取得

の 2 つを提案している．実験では，それぞれのベースラインとして

- (WN)K のみで Web 検索を行い，検索結果のスニペット集合から出現回数の多い語を C の要素として取得
- (YN)K のみで知恵袋検索を行い，検索結果の QA コンテンツの回答集合において出現回数の多い語を C の要素として取得

という C を取得する 2 つの方法を追加し，それぞれの手法で得られた C を用いて $Score_{pn}$ を計算する．

P は K を Web 検索エンジンに用いて得られた検索結果の上位 10 件のページとする．各ページについてそれぞれの手法で $Score_{pn}$ を計算し，$Score_{pn}$ の値の高い順に P の要素を並び替える．$Score_{pn}$ が同値のページは検索エンジンの出力順位に従って並び替えを行う．

この操作を 10 個の QA コンテンツに対して行い，それぞれの結果について「質問 q に対する回答を含むページ」を正解ページとして再現率と適合率を求める．QA コンテンツは回答数が 3 つであり，かつ別解の存在する質問内容であるものを人手で集めたものとする．再現率を 0 から 0.1 刻みで増加したときの補間適合率の平均を求め，11 点平均適合率を計算した．結果を図 1 に示す．

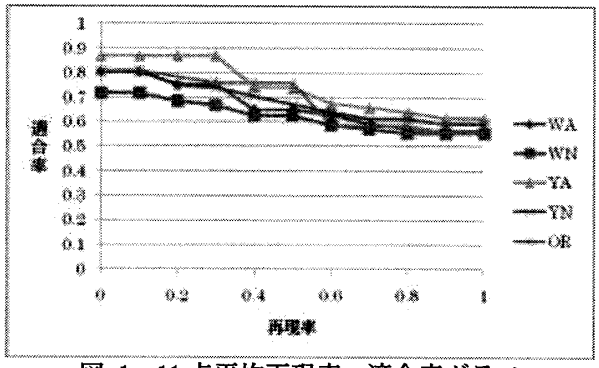

図 1 11 点平均再現率・適合率グラフ

図 1 において，OR は検索エンジンの元の結果の再現率・適合率を計算したものであり，本評価実験のベースラインとなる．

図 1 より，OR を上回ったのは(YA)手法のみという結果となった．提案手法の精度が全体的に高くない原因として，$Score_{pn}$ の正規化を行っておらず，語が多く含まれるページが必然的にスコアが高くなったということが考えられる．

(WN)，(YN)より(WA)，(YA)の精度が高いことから，回答に共通して現れる語を取得する際に対象としている QA コンテンツの回答情報を用いることは有効であるといえる．

4. まとめと今後の課題

本稿ではある QA コンテンツの質問 q をもとにしたクエリで収集した Web ページ集合の各ページについて，そのページが q に対する回答情報を含むかどうかを判断するスコアの計算方法について提案した．　今後は，スコアの計算時に重要となる，q に対する回答表現に共通して現れる語を精度よく取得する方法や，本研究の最終的な目標である QA コンテンツに対する別解情報取得手法について考えていく予定である．

謝辞

本研究の一部は，京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」，および，科学研究費補助金（課題番号 18049041，21700105），および，NICT 委託研究「電気通信サービスにおける情報信憑性検証技術に関する研究開発」によるものです．ここに記して謝意を表します．

参考文献

[1]佐藤,石下,森：“Web 文書を情報源とする記述的な回答が可能な応答システム”,言語処理学会第 14 回年次大会発表論文集,pp.1009-1012 (2008)

[2]高田,山本,小山,田中：“質問応答コンテンツに対する Web からの別解情報検索”, 第 2 回知識共有コミュニティ WS 論文集,pp.53-62(2009)

[3]馬,田中：“話題構造に基づく放送と Web コンテンツの統合のための検索機構”,情報処理学会論文誌,45,pp.18-36(2004)